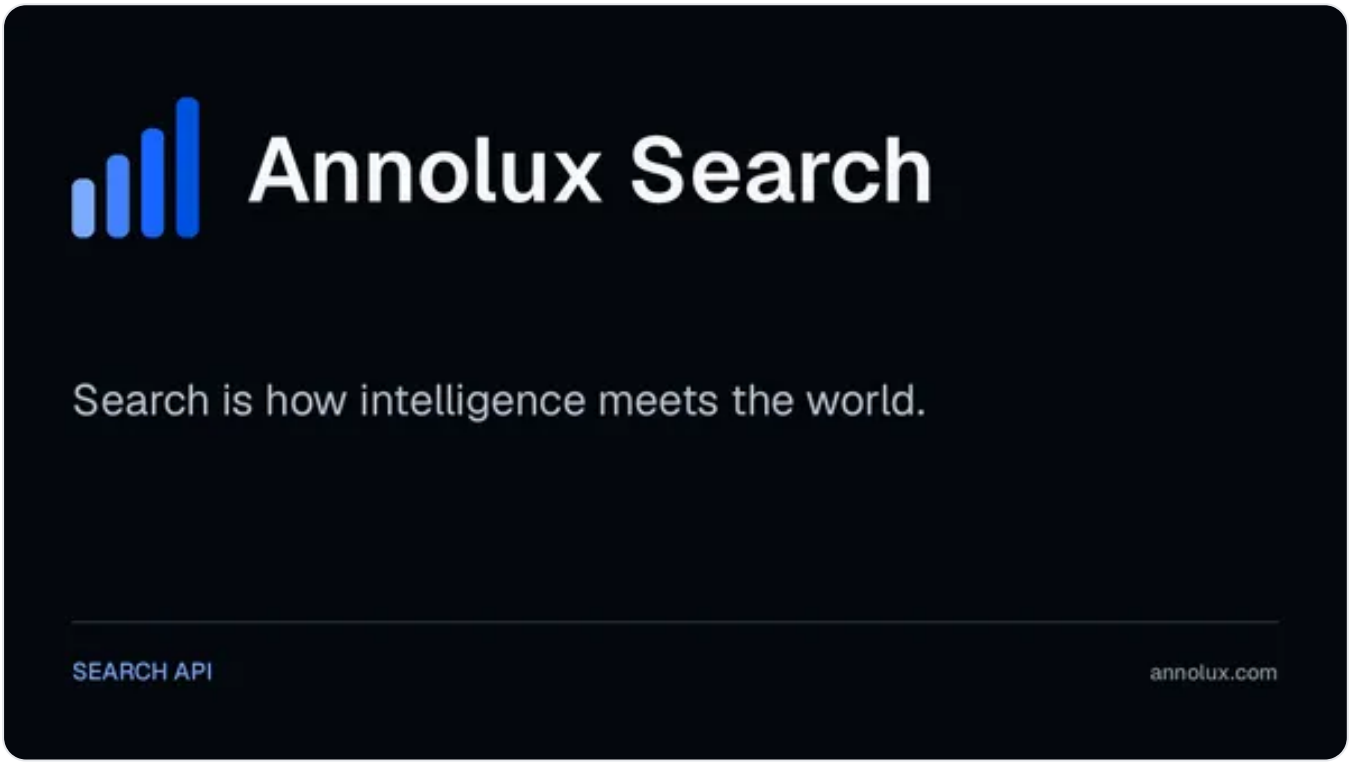
Annolux

A bilingual search API for AI agents and RAG workflows with source links, fetched timestamps, and predictable credit usage.



About

Annolux is a bilingual search API for AI agents and RAG systems. It is built for English and Chinese retrieval, so agents can search a curated index and cite results with source URLs and fetched_at timestamps. The service emphasizes clean snippets, deduplication, and deterministic credit billing, which makes usage easier to track in automated workflows. Annolux also exposes MCP support and public benchmarks for users who want to evaluate retrieval quality before integrating it.

What is Annolux

Annolux is a bilingual search API that serves English and Chinese retrieval use cases for

general-purpose web browser, and the product materials say the index contains 337,000+ pages that are owned and maintained by the team. Each query returns source-attributed snippets and a fetched_at timestamp so downstream systems can show where an answer came from and when the page was read.

The service is designed around single-call retrieval. Instead of asking an agent to search, fetch, and normalize separately, Annolux returns structured results from its API in one step. The example response shown on the site includes a title, URL, snippet, rank, and fetched_at value. The docs also mention that results are deduplicated.

Key features

- English and Chinese search coverage for bilingual applications.
- Source URL and fetched_at included with each result for citation-aware workflows.
- Clean markdown-style snippets with boilerplate stripped from fetched pages.
- Deduplicated results to reduce repeated sources in ranked output.
- API responses shown with explicit credit usage headers such as X-Annolux-Credits-Used and X-Annolux-Credits-Remaining.
- MCP-ready access, which suggests the search service can be wired into agent tooling that understands the Model Context Protocol.
- Public benchmark reporting, including held-out evaluation cases, HIT@10, MRR@10, and p95 latency figures.
- A documented rate limit of 5 RPS with burst 10, based on the site's example and messaging.

How to use it

- Sign in and create an API key from the Annolux site.
- Read the docs for the search endpoint and request format.
- Send a POST request to the search API with a query and optional limit.
- Parse the returned JSON fields, especially title, url, snippet, rank, and fetched_at.
- Use the source URL and timestamp in your agent, citation layer, or RAG prompt assembly.

A typical use case is asking for a topic such as “rust ownership” and then feeding the returned snippets into a retrieval-augmented generation pipeline. The site’s example also shows that the same result can be consumed by an agent without needing an extra fetch step.

Who it is for

Annolux is a good fit for teams building AI agents, search-backed assistants, or multilingual RAG systems. It is especially relevant if your product needs:

- Citations that point to the original source page.
- Search results in both English and Chinese.
- A retrieval service with predictable per-call credit accounting.
- A smaller, purpose-built index instead of a general web search product.
- Tooling that can be connected to an agent framework or MCP-compatible stack.

The product may also suit developers who want to benchmark retrieval quality before production rollout, since the site highlights frozen blind evaluation and public score reporting.

What to know before you use it

Annolux is not presented as a full web search engine, crawler dashboard, or document ingestion platform; it is a search API for retrieval. The public materials do not list every supported parameter, authentication detail, or plan tier, so you should rely on the official docs and pricing pages for current limits and implementation specifics.

The benchmark numbers on the site are tied to a frozen blind set with 40 held-out queries, so they are useful for context but not a guarantee for every workload. Latency, rank quality, and snippet relevance can vary by query, language, and domain. The index size is described as 337,000+ pages, which is substantial but still narrower than the open web, so niche or newly published pages may not appear immediately.

Because the service returns fetched_at timestamps, it can help you tell when a page was read, but that does not mean the underlying page has not changed since then. For time-sensitive or compliance-sensitive workflows, verify critical facts directly on the source

When you need a bilingual search API for AI agents and RAG, Annolux is built to provide citeable results with timestamps, deduplication, and credit-based accounting in one request.

FAQ

How does Annolux bill usage?

Does Annolux support both English and Chinese search?

Can I use Annolux in an AI agent or MCP workflow?

What data comes back in a search result?

Is there a public benchmark for Annolux?

Sponsored

Belay – Awa... Featured
Developer Tools

Stays awake while your agents work.

Free # Developer Tools

View ↗

Partly Featured
Finance

Partly is a mobile app that lets friends quickly scan receipts, split any bill fairl...

Free # Finance

View ↗

Visit website ↗

annolux

Category	AI
Platform	Web, API, MCP
Launched	Aug 2026
Updated	today
Verified	yesterday

TAGS

search api # ai-agents # bilingual search # citations # rag # credits # mcp # web search

© 2026 Dir Hub

Annolux